3 October 1969

MEMORANDUM FOR: Chief, Language School

THROUGH : Director of Medical Services
Director of Training

SUBJECT : Language Validation Report

1. Attached PSS/OMS Research Report describes a recent validation study of the Artificial Language Test, a general test of aptitude for learning a foreign language. Agency employees who were in language training for some period during FY68-69 and on whom aptitude and objective achievement ratings were available constituted the group studied.

2. The finding, in brief, was that a significant degree of relationship does exist between the Artificial Language Test and training achievement. However, the size of the relationship is not so large as to recommend the exclusive use of the test in decisions for which the potential for language acquisition is a major factor.

3. The Psychological Services Staff appreciated the opportunity to collaborate with the Language School in this research approach to a problem of mutual concern, and we want to thank the members of the Language School Staff who contributed so substantially to the conduct of the study.

25X1A

[redacted]

Acting Chief, Psychological Services Staff
Office of Medical Services

RESEARCH REPORT

A STUDY OF SOME CORRELATES OF THE ARTIFICAL LANGUAGE TEST

[redacted] 25X1A

Psychological Services Staff
Office of Medical Services
September, 1969

Introduction

This research report describes a recent validation study of the Artificial Language Test (AL-AGO), a general test of aptitude for learning a foreign language. The test is presently included in the morning portion of PATB, the battery of tests given by the Psychological Services Staff to applicants for professional positions in the Agency. AL-AGO has been part of PATB since December 1956.

Between February 1952, and October 1965, the AL-AGO had been used in combination with other measures of language aptitude to predict success in Agency language courses. In the intervening years a number of in-house validity studies had been made on all of these language aptitude tests. Typically, it was found that AL-AGO provided the best single predictor of achievement in Agency language training. Reports of validity coefficients in the .40's and .50's were not at all unusual. Review of this early research literature suggested that AL-AGO served as a highly satisfactory aptitude measure. In light of the very high correlations - in the .70's - between AL-AGO and overall Foreign Service Institute (MLAT) ratings, it is not surprising that AL-AGO should have had the degree of relationship it did with achievement in language training. Recently a State Department memorandum reported a correlation in the .60's between MLAT ratings and language course evaluations for a group of 691 trainees.

As of FY70 all candidates for Agency language training have been required to take the full MLAT. Consequently, AL-AGO is now treated as only an initial and approximate measure of an applicant's general language learning aptitude. It is felt by the Language School that the multi-score MLAT is the more appropriate source of information for differential placement, training emphases, student management, and selection in special cases.

Objectives

Good psychometric practice recommends that aptitude tests be periodically revalidated. This is especially true if changes have occurred in the student population, the manner in which training is conducted, course objectives, evaluations, etc. With this in mind, PSS in cooperation with the Language School undertook a study of the relationship between AL-AGO ratings and rate of achievement in language training. A secondary objective was to compare the distribution of AL-AGO ratings of a representative group of language trainees with that of a professional applicant sample. Results of this comparison would indicate if these aptitude ratings had been systematically used for screening candidates for language training.

Procedure

The starting point of this study was the preparation of a listing of all individuals who studied a foreign language for some period of time during FY68 or 69 and on whom at least

some language evaluations were recorded. FY68 and 69 were selected because language evaluations from these years were reasonably accessible and were administered in a relatively standardized way. Prior to these years, there was some question as to the format followed and the degree of standardization in assigning language evaluations.

The listing of students was organized by fiscal year within each language. For each individual the following pieces of information were recorded:

Bases of evaluation:

Student proficiency levels both at the outset of the training and at its termination were based on either the instructor's estimate or on an "official" test. A coding procedure was devised to differentiate between initial and final rating bases. It was felt that relying on an official test for determining proficiency level would minimize the amount of subjectivity, and probably the error variance, in the ratings. Hence, only those final ratings based on an official test were subjected to analysis. For two reasons this restriction was not applied to entry level proficiency ratings. First, doing so would have cut the study sample down to too small a size. Secondly, it was felt that where a student had no previous knowledge of the language he was to study (this was very typically the case) his initial level rating would not differ whether assigned by the instructor's estimate or on the basis of a test.

Language elements rated:

The Language School training evaluation form typically recorded student achievement in four elements of the language studied: reading, speaking, pronunciation, and understanding. For the purposes of this study, only the reading and speaking elements were considered. The pronunciation element was omitted on the assumption that it had minimal rational relationship to the design and intent of the AL test. The understanding element was omitted on the assumption that the requirements for displaying competency on a written test, like AL, and the aural aspect of understanding the spoken words were quite different. The reading element was included because of the written nature of AL, and speaking because of the similarity

between actual speaking and the reconstruction of words and forms required by the AL test.

Languages:

Twenty languages were represented in the overall listing of students who had been in language training some time during FY68-69. The distribution of students in these different languages is shown in Table I. It can be seen that the majority of the languages studied, 13 in fact, had fewer than ten enrollees; only seven languages had more than ten students who received final proficiency levels based on an official test.

On the advice of the Language School, ratings of students in Vietnamese and Thai were treated together. This was true for German and Swedish ratings also. Students of Russian--there were only 17 who had received their final proficiency level by official test--were not included in our analyses because many were in courses designed to familiarize them with the Cyrillian alphabet for filing purposes. Hence, ratings in reading and speaking were not altogether appropriate criteria for this group.

Gain:

Since knowledge of a student's final proficiency level without knowledge of his entry level did not give any indication of his progress in language training, the basic datum used in this study consisted of the gain or final minus initial level registered by each student.

During the period from which language evaluations were drawn for this study, minor changes were made in the rating of student achievement. Specifically, "pluses" indicating particularly noteworthy achievement at a given level were assigned to the five basic proficiency levels (No Proficiency, Slight, etc.). Since 1 February 1969 pluses were given in all course elements. Between 14 November 1968 and 1 February 1969 students could receive pluses in all elements except reading and prior to 14 November 1968 pluses were given only in the speaking element of the language course. To maintain uniformity in assessing student achievement and greatly simplify the data analysis, we have chosen to disregard plus marks for this study.

Hours in Study:

For each student both nominal and actual hours of language training were recorded. Nominal hours are the number of hours a student was assigned to the Language School. Actual hours are the number of hours spent in actual study. For obvious reasons, actual hours in combination with gain were used exclusively in describing a student's achievement.

S-E-C-R-E-T

TABLE I

DISTRIBUTION OF STUDENTS[1] BY LANGUAGE STUDIED AND BY BASIS OF ASSIGNMENT OF INITIAL AND FINAL PROFICIENCY LEVELS

BASIS OF RATING 1=Initial; 2=Final	ARABIC	CHINESE	CZECH	FRENCH	GERMAN	GREEK	INDONESIAN	ITALIAN	JAPANESE	PERSIAN	POLISH	PORTUGUESE	ROMANIAN	RUSSIAN	SERBO-CROATIAN	SPANISH	SWEDISH	THAI	TURKISH	VIETNAMESE	
1. Instructor's Estimate (I.E.) 2. Official Test (O.T.)	0	7	2	58	23	3	7	8	3	1	1	6	1	14	3	38	2	9	5	22	213
1. O.T. 2. O.T.	0	8	0	14	8	0	0	1	2	0	0	1	0	3	0	14	1	1	1	1	55
1. I.E. 2. I.E.	0	2	0	3	3	4	0	0	0	3	0	1	0	36	0	4	0	1	0	1	58
1. O.T. 2. I.E.	0	1	0	2	0	0	0	0	0	0	0	0	0	4	0	3	0	0	0	0	10
TOTALS	0	18	2	77	34	7	7	9	5	4	1	8	1	57	3	59	3	11	6	34	336

[1] Only students on whom AL scores and both initial and final speaking proficiency levels were available are included in this breakdown.

S-E-C-R-E-T

AL ratings:

Adjectival ratings and actual raw scores on the AL test were recorded where available. Raw scores on AL for the study sample ranged from 5 to 57---59 is a perfect score. Some checks were made to determine if using raw scores rather than the five adjectival ratings would lead to greater predictability of rate of achievement. There was no evidence that it did so, and hence the simpler adjectival ratings were used in the analyses described below.

Although 423 individuals were identified as having studied a language during FY68-69, complete data were available on only 351 cases. About 250 of these individuals had received their final proficiency levels by official test and hence were eligible for inclusion in our subsequent analyses.

Results

I. DISTRIBUTION OF AL RATINGS

In the first part of this result section, we compare Agency language trainees and applicants on performance on the AL test. If the AL test has in the past systematically played a part in screening candidates for language training, we would expect that proportionately more of the trainees than applicants had received ratings of Good and Very Good and fewer Poors and Very Poors. Marked restrictions in the distribution of AL ratings in the validation sample arising from pre-selection on AL would spuriously lower any relationship between aptitude and achievement ratings.

To check for restriction of range, the distribution of AL ratings of 3,282 men and women applicants tested in FY67 for professional positions was compared with that of the 351 language trainees. We had no reason to believe that these

two groups differed, on the average, in age and educational level attained. The percent of each group receiving the various aptitute ratings is shown below.

AL RATING

	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD	
Professional Applicants	31	37	32	$x^2=5.94$, df=2 $P > 05 < 10$
Language Trainees	25	38	37	

The differences are modest in size and just fail to reach the commonly accepted level for claiming statistical significance. They seem to indicate that AL scores have, only to a very limited degree, been involved in the enrollment decision.

In the above comparisons, the language trainee group consisted of all individuals on whom AL and achievement ratings were available. Since not all of these students were included in the analyses relating aptitude and achievement ratings, an additional comparison involving only the actual validation sample consisting of 148 individuals and the professional applicant group seemed in order. Results are shown below.

AL RATING

	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD	
Professional Applicants	31	37	32	$x^2=2.54$, df=2 not significant
Validation Sample	27	43	30	

It appears that the validation sample is not to any significant degree pre-selected on AL scores. Although this result indicates that obtained relationships between aptitude and achievement have not been lowered by restriction of range in the sample, it raises questions regrading the use to which the AL test is put. A thorough examination of this problem with the data available would go beyond the objectives of the present research.

A final analysis on AL ratings involved comparing their distribution among students in differing language families. It was felt that the esoteric languages (eg., Vietnamese, Chinese, Russian) placed greater demands on students than the world languages (eg., Spanish, French, Italian). Hence, we questioned whether individuals with higher language aptitude were selected (or selected themselves) for training in the more demanding languages. Results of these comparisons indicated that only to a slight degree did average aptitude ratings differ among languages. The differences were in the expected direction ($F=2.94$, $df=4, 311$, $P<.05$). The mean aptitude level of students in the Romance and Germanic language families fell at about the "average" point. The mean level for students of Vietnamese and Thai was very slightly above this and the mean level for students in the Oriental and Slavic languages still higher--actually somewhere between Average and Good on the rating scale.

II. RELATIONSHIP BETWEEN AL AND ACHIEVEMENT

On the basis of conversations with the Language School, it was decided not to treat gains in proficiency levels as if they fell on an equal-interval scale. That is, the five-point proficiency scale ranging from (1) No Proficiency to (2) Slight to (3) Elementary to (4) Intermediate to (5) High, was not assumed to consist of equi-distant points. The student moving from No Proficiency to Elementary, a two-step gain, did not necessarily achieve to the identical degree as one who moved from Slight to Intermediate, also a two-step gain. Nor was it assumed that a gain, say from Slight to Elementary in one language, represented the same degree of accomplishment as an identical gain in another language. These considerations dicatated a rather specific type of analysis to ascertain relationships between language aptitude and rate of achievement in language training. The analysis is described below.

For each language studied all gains realized by students from entry to final level were listed (eg., Slight to Intermediate, No Proficiency to Elementary) separately for the reading and speaking elements of the course. Those students not realizing any gain in rated proficiency were not included in the analyses because of the problems in interpretation their cases would present. Within each gain category, the number of actual hours spent by students in language training was arranged from low to high. Each student's actual hour entry was accompanied by his AL adjectival rating. The distribution of actual hours was split at the median for each gain level, separately for each

language. The resulting bifurcated distributions were arranged into summary tables as in the model shown below--one for reading proficiency, one for speaking.

AL RATING

GAIN ACCOMPLISHED IN:	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD
ABOVE MEDIAN NUMBER OF HOURS			
BELOW MEDIAN NUMBER OF HOURS			

A tally in Row One, Column One, would indicate that the student had received a low aptitude rating and had achieved his specific gain in proficiency in more than the median number of hours required by other students in that same course realizing that same gain. For ease in interpretation all entries have been converted to percents with the percents in each column totaling 100.

Tables 2A and B and 3A and B present the basic data from the analysis described above. Table 2A shows the contingency between aptitude and rate of achievement (defined above or below median number of hours) on the reading element for the Romance Languages. This sub-sample consisted of 62 French, 36 Spanish, four Italian, and two Portuguese language trainees. Table 2B shows the contingency between aptitude and rate of achievement on the speaking element for students in these languages. Both of these tables indicate that of those students

TABLE 2A

ROMANCE LANGUAGES (N=104) READING

AL RATING

GAIN ACCOMPLISHED IN:	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD
Above Median Number of Hours	61%	52%	35%
Below Median Number of Hours	39%	48%	65%

TABLE 2B

ROMANCE LANGUAGES SPEAKING

	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD
Above Median	70%	43%	40%
Below Median	30%	57%	60%

TABLE 3A

COMBINED LANGUAGES (N=148) READING

	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD
Above Median	62%	56%	32%
Below Median	38%	44%	68%

TABLE 3B

COMBINED LANGUAGES SPEAKING

	POOR, VERY POOR	AVERAGE	GOOD, VERY GOOD
Above Median	70%	47%	39%
Below Median	30%	53%	61%

who received ratings of either Good or Very Good on AL, the clear majority achieved their particular proficiency gains in less than the median number of hours (65 vs. 35 in reading, 60 vs. 40 in speaking). Conversely, more students with aptitude ratings of Poor and Very Poor take above the median number of hours to achieve their particular gains (61 vs. 39 in reading, 70 vs. 30 in speaking).

Tables 3A and B show the above relationships for several language groups combined. To the data of students in Romance Languages has been added that of some 18 students in German and Swedish and 30 students in Vietnamese and Thai. Essentially the same sort and degree of relationship obtained for this slightly larger (n=148) group of language trainees. The differences in distributions in rate of achievement across aptitude categories are statistically reliable. On the basis of these findings, we concluded that aptitude ratings are positively associated with rate of achievement in the languages represented in the sample. The degree of association, although statistically reliable, is, in an absolute sense, a modest one. Caution in interpreting AL rating is advised.

To give some indication of the differential number of hours required for a particular proficiency gain among students receiving low and high AL ratings, we have looked at some selected data from the Romance Language group. On the reading element, those students with low aptitude ratings (Very Poor and Poor) took, on the average, 581 hours to realize a gain from

No Proficiency to Elementary. Students with high aptitude ratings (Good and Very Good) took, on the average, 459 hours to realize this same gain. This difference in average number of hours fails to reach commonly accepted levels of statistical significance. On the speaking element of the Romance Languages, the low aptitude group went from No Proficiency to Elementary in 720 hours, on the average. For the high aptitude group, this gain was made in 537 hours. This difference is statistically significant.

Conclusion

It may be concluded that among the languages under study, primarily the Romance Language family, a significant degree of relationship exists between AL ratings and a derived index of rate of achievement in actual language training. However, in absolute terms, the size of this relationship is not so large as to recommend the exclusive use of AL ratings in decisions for which the potential for language acquisition is a major factor. Available evidence would suggest that the individual's practical and academic experiences with languages, his scores on the MLAT, his motivation, etc., should be considered along with his AL rating in arriving at any decision regarding language training.